Relatório descritivo dos dados em sala de aula da amostra perfeita, 10p e 5p

Thiago Félix Teixeira Chaves

[1]Instituto de Educação Superior de Brasília (IESB)

thiago.chaves@iesb.com.br

Resumo. *Uma análise da notas de variáveis Idade, Nota de matemática, ciências da natureza, linguagens e códigos, ciências humanas e redação*

1. Introdução

Este relatório tem como objetivo apresentar uma análise descritiva das amostras P10, P5 e da amostra perfeita dos dados do **Exame Nacional do Ensino Médio (ENEM) 2023**. Os dados incluem informações sobre os candidatos, suas performances e notas nas disciplinas propostas pelo exame. As variáveis selecionadas para esta análise são: REDAÇÃO, MATEMÁTICA, LINGUAGENS, CÓDIGOS, CIÊNCIAS DA NATUREZA, CIÊNCIAS HUMANAS e a idade dos candidatos.

O Exame Nacional do Ensino Médio (ENEM) de 2023 teve início no dia 5 de novembro e continuou com a segunda etapa no dia 12 do mesmo mês. As provas seguiram o cronograma abaixo:

1º dia - 05/11:

- Redação
- 45 questões de Linguagens e Códigos
- 45 questões de Ciências Humanas

2º dia - 12/11:

- 45 questões de Ciências da Natureza
- 45 questões de Matemática

Em 2023, o ENEM registrou aproximadamente 3,9 milhões de inscritos, um aumento de 13,1% em relação a 2022. Para este relatório, foram extraídas três amostras do total de candidatos:

- **Amostra Perfeita**: 53.427 candidatos
- **Amostra de 5%**: 104.799 candidatos
- **Amostra de 10%**: 209.598 candidatos

Por meio da visualização de medidas estatísticas, será possível observar como essas informações se distribuem, avaliando se a maioria dos candidatos obteve boas notas e qual é o perfil do público que participa do ENEM. Além disso, incluiremos gráficos de histogramas e Boxplot, que facilitarão a compreensão dos dados apresentados.

O intuito dessas informações estatísticas e gráficas é proporcionar a educadores e instituições uma visão clara sobre o perfil dos participantes do exame, que é de grande interesse para esses profissionais. A análise, em certos aspectos, preserva a anonimidade do público para aprofundar a discussão sobre os dados coletados. O objetivo final é auxiliar na tomada de decisões que possam levar a melhorias no ensino.

2. Objetivos

Analisar as variáveis de REDAÇÃO, MATEMÁTICA, LINGUAGENS, CÓDIGOS, CIÊNCIAS DA NATUREZA, CIÊNCIAS HUMANAS e IDADE dos **candidatos do Exame Nacional do Ensino Médio (ENEM) 2023**, utilizando gráficos e medidas de resumo para facilitar a compreensão das informações.

2.1. Objetivos gerais

- Apresentar Medidas resumos(medidas estatísticas).
- Apresentar gráficos da distribuição dos gráficos(Histogramas e Boxplot).
- Explicação geral dos gráficos e medidas resumos apresentadas.

3. Fundamentação Teórica

Análise da variável Idade

Ao observar as medidas de resumo da variável idade nas três amostras, podemos verificar que os dados são bastante semelhantes e próximos, indicando que as amostras representam bem a população.

Analisando as três amostras, percebemos que a média varia entre 20 e 22 anos, resultando em um desvio padrão semelhante, em torno de 6. O coeficiente de variação fica próximo de 30% a 31%. A principal diferença entre as amostras é a quantidade de dados analisados: 104.799, 209.598 e 534.527 para as amostras da P5, P10 e amostra completa, respectivamente.

Podemos resumir a análise geral da seguinte forma:

- 50% dos candidatos têm menos de 18 anos, e 75% têm menos de 24 anos, sugerindo que a maioria é formada por jovens.
- Embora exista uma minoria de idades mais avançadas, podendo chegar a 70 anos (a idade máxima), essa faixa não está claramente representada.
- As idades estão bem concentradas em torno da média, como indicado pelo desvio padrão relativamente baixo de 6,99 e uma média em torno de 19 anos. Ou seja, as idades não se afastam muito da média.

Por fim, embora os dados sejam consistentes, não podemos considerá-los homogêneos, e sim heterogêneos. Isso se deve ao fato de que, em todas as tabelas, o coeficiente de variação está acima de 30%, ainda que por uma pequena margem.

Ao olharmos para os gráficos Boxplot, podemos verificar que todos eles tendem a outliers (valores distantes da média) de valores elevados, isto por que mesmo que a maior parte dos canditados estajam entre 20 anos de acordo com o histograma, não podemos esquecer que existem os demais que podem a chegar a candidatos com idade de 70 anos, por esse motivo, podemos verificar essas informações junto ao boxplot. Isso nos dar a entender que esses valores não são o comum a acontecer, pois se distancia do publico em massa, mas para análises especificas, são dados valiosos que nos mostram que o exame tem alcançado pessoas de idade mais avançadas.

Analysis Variable : Idade													
Mean	Std Dev	Minimum	Maximum	Median	N	N Miss	Std Error	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
22.0979654	6.9931772	16.0000000	70.0000000	19.0000000	53427	0	0.0302548	48.9045268	18.0000000	54.0000000	31.6462490	18.0000000	24.0000000

Figura 1. Tabela da amostragem Perfeita

Analysis Variable : Idade												
Mean	Std Dev	Minimum	Maximum	Median	N	Std Error	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
20.9005811	6.2972932	16.0000000	70.0000000	18.0000000	104799	0.0194525	39.6559014	18.0000000	54.0000000	30.1297516	18.0000000	21.0000000

Figura 2. Tabela da amostragem P5

Analysis Variable : Idade												
Mean	Std Dev	Minimum	Maximum	Median	N	Std Error	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
20.9101136	6.2865144	16.0000000	70.0000000	19.0000000	209598	0.0137314	39.5202629	18.0000000	54.0000000	30.0644677	18.0000000	21.0000000

Figura 3. Tabela da amostragem P10

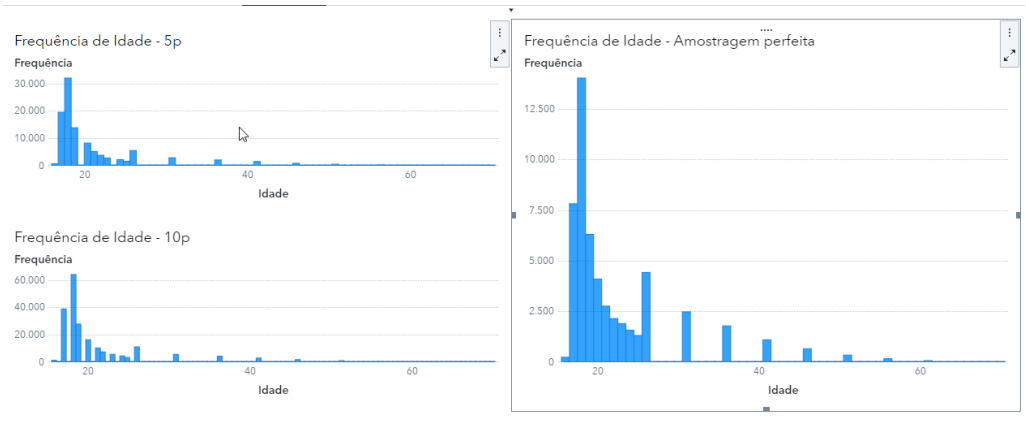

Figura 4. Histograma da IDADE

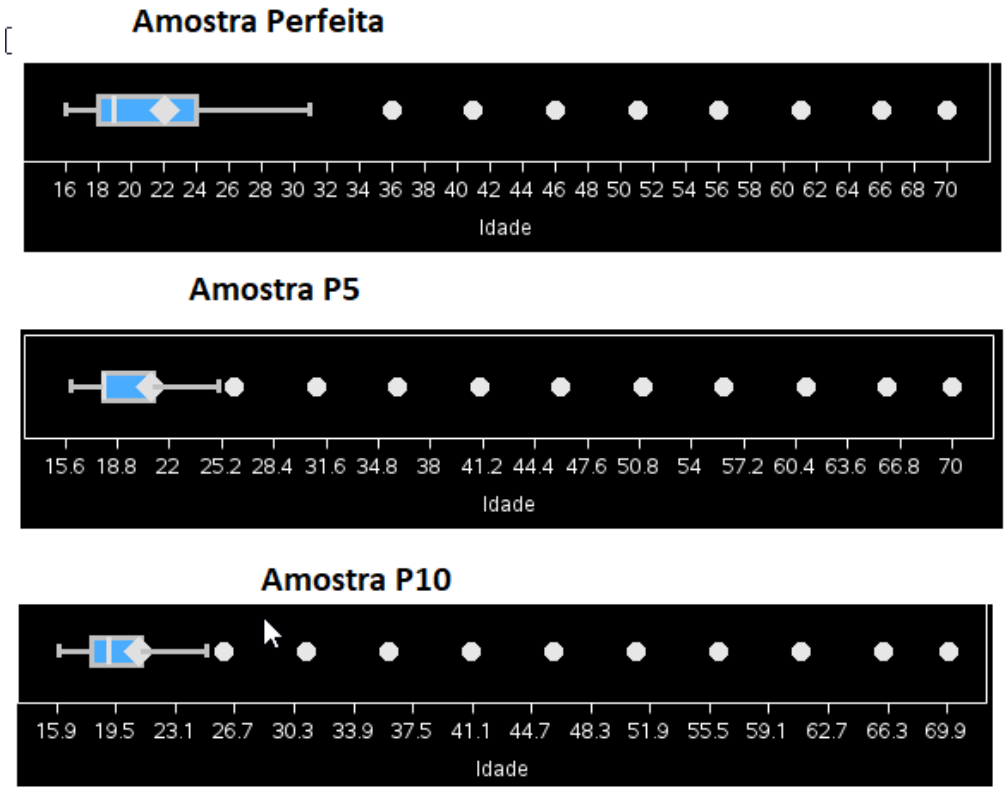

Figura 5. Box-Plot da IDADE

Análise da variável de Matemática

Ao observar as tabelas de amostragem da nota de matemática, vemos que o P5 e o P10 são bem semelhantes em todas as medidas de resumo. No entanto, na **amostra perfeita**, alguns valores se distanciam significativamente das outras duas. Essa diferença fica clara no gráfico de histograma, onde as amostragens de 5% e 10% mostram uma **assimetria à direita** — ou seja, há uma grande quantidade de dados acumulados à esquerda, que vão diminuindo conforme se deslocam para a direita. A explicação para isso será discutida logo abaixo.

Já no histograma da **amostra perfeita**, observamos uma **distribuição simétrica**: os dados se acumulam no meio, com uma distribuição equilibrada tanto para a esquerda quanto para a direita. Isso também será explicado melhor em seguida.

É interessante destacar que, quanto menor a amostra, piores tendem a ser os resultados, e a **amostra perfeita** é a menor entre todas, com 53.427 dados, seguida de 104.799 para o P5 e 209.598 para o P10. É algo a se pensar!

Ao observarmos os histogramas, podemos entender a distribuição dos dados através das medidas de resumo. Para isso, temos a seguinte regra:

- **Gráfico simétrico**: quando os dados estão bem distribuídos para os dois lados (esquerda e direita), temos **Média = Mediana = Moda**.

- **Assimetria à esquerda**: quando a maioria dos dados está à direita e os valores vão diminuindo para a esquerda, criando uma "cauda"no gráfico. Nesse caso, **Média < Mediana < Moda**.
- **Assimetria à direita**: quando a maioria dos dados está à esquerda e os valores vão diminuindo à medida que aumentam para a direita, também criando uma "cauda". Aqui, **Moda < Mediana < Média**.

Agora, ao analisar os gráficos:

- O gráfico da **amostra perfeita** mostra uma **distribuição simétrica**, pois a média, mediana e moda estão muito próximas: 535,65, 540,90 e 546,30, respectivamente.
- Nos outros dois gráficos (P5 e P10), as medidas de resumo indicam uma **assimetria à direita**, ou seja, a moda é menor que a mediana, que por sua vez é menor que a média. Para o P5, esses valores são: 0, 524,70 e 536,43.

Algo interessante é que as médias das três amostras são parecidas: 535,65 (amostra perfeita), 535,88 (P10) e 536,43 (P5). Mas a atenção deve estar no **desvio padrão (STD DEV)**, onde vemos uma diferença significativa: para as amostras P10 e P5, o desvio padrão é alto (131,55), enquanto na amostra perfeita é bem menor (66,43). Isso é importante porque a média é uma medida muito sensível a **outliers** (valores muito altos ou muito baixos), o que pode distorcer a análise. O desvio padrão ajuda a entender o quanto os dados estão espalhados em relação à média: quanto maior o desvio padrão, mais dispersos estão os dados.

Mesmo com desvios padrões diferentes, as médias ficaram próximas. Isso levanta a questão de por que isso acontece, já que a dispersão dos dados é bem distinta.

Ao analisarmos os gráficos de boxplot das amostras p10 e p5, observamos informações interessantes sobre a distribuição das notas. Primeiramente, notamos a presença de *outliers* – valores que se distanciam da média e estão fora do retângulo em formato de cristal no gráfico, que representa o intervalo interquartil (IQR) onde a maior parte dos dados se concentra.

Tanto à direita quanto à esquerda da média, identificamos *outliers*, o que indica a presença de notas anormalmente altas e baixas. Essas notas variam desde 0 até 958, conforme indicado na amostra p10. Isso significa que há estudantes que tiraram notas extremamente baixas, enquanto outros atingiram notas muito elevadas, demonstrando uma dispersão significativa entre as pontuações.

No gráfico da **amostra perfeita**, vemos uma distribuição semelhante, com uma diferença: os valores mínimos não chegam a extremos como 0, pois o menor valor é 328,20. Isso sugere uma distribuição um pouco mais concentrada, especialmente em comparação com p5 e p10,.

Em resumo:

- A **amostra perfeita** apresenta uma distribuição mais equilibrada, com menor variabilidade nas notas em comparação com as outras duas amostras. Contudo, devido ao menor tamanho da amostra, há uma limitação na confiabilidade dos resultados.
- As amostras **p10** e **p5**, apesar de terem tamanhos diferentes, fornecem resultados muito semelhantes, o que aumenta a confiança de que esses dados representam melhor a população.

Além disso, observamos que a maioria das notas está concentrada abaixo de 600 pontos, com poucos estudantes alcançando notas acima dessa faixa. Esses valores altos são considerados *outliers* no sentido estatístico, mas, no contexto do ENEM, são justamente esses valores que recebem mais atenção, pois representam os candidatos de melhor desempenho – um dos focos de avaliação no exame.

Analysis Variable : NOTA_MATEMATICA Nota da prova de Matemática												
Mean	Std Dev	Minimum	Maximum	Median	N	Std Error	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
535.6552137	66.4320942	328.2000000	822.1000000	540.9000000	53427	0.2874071	4413.22	546.3000000	493.9000000	12.4020251	493.7000000	581.3000000

Figura 6. Tabela da amostragem Perfeita

Analysis Variable : NOTA_MATEMATICA Nota da prova de Matemática												
Mean	Std Dev	Minimum	Maximum	Median	N	Std Error	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
536.4331005	131.5537856	0	958.6000000	524.7000000	104799	0.4063729	17306.40	0	958.6000000	24.5238009	432.7000000	632.6000000

Figura 7. Tabela da amostragem P5

Analysis Variable : NOTA_MATEMATICA Nota da prova de Matemática												
Mean	Std Dev	Minimum	Maximum	Median	N	Std Error	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
535.8836935	131.0525395	0	958.6000000	524.2000000	209598	0.2862542	17174.77	0	958.6000000	24.4554072	432.4000000	631.6000000

Figura 8. Tabela da amostragem P10

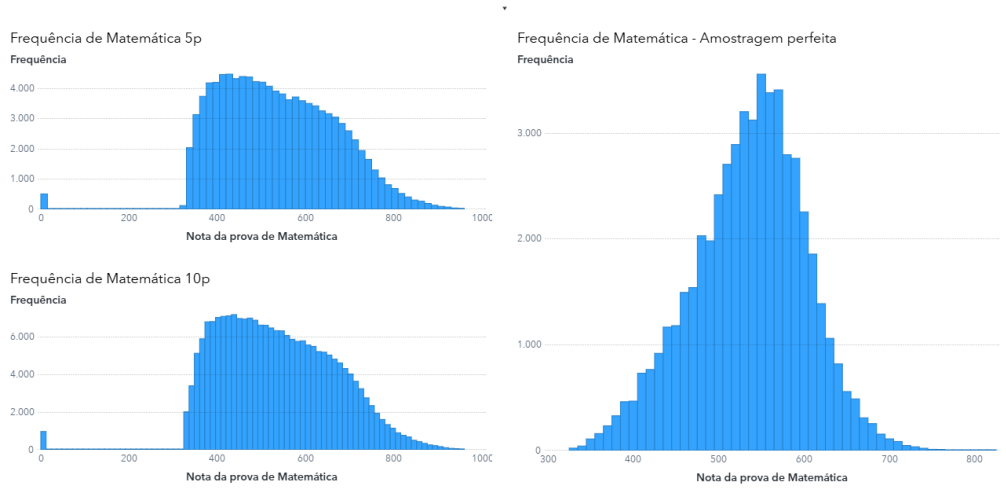

Figura 9. Histograma da Matemática

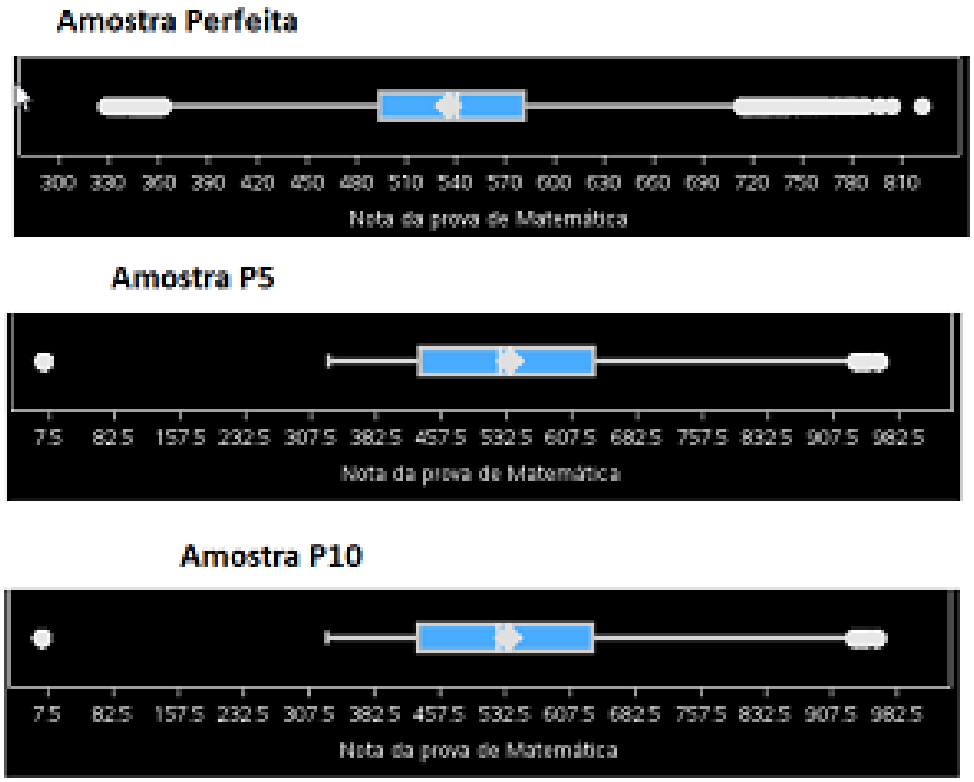

Figura 10. Box-Plot da Matemática

Análise da variável de Linguagens e Códigos

Analisando a variável "Linguagens e Códigos"nas amostras de 10%, 5% e na amostra perfeita, observa-se que todas as medidas-resumo são bastante próximas entre si, a ponto de serem quase idênticas, mesmo com quantidades diferentes de registros em cada amostra: 209.598, 104.799 e 53.427, respectivamente. Essa semelhança é confirmada pelo formato dos gráficos, que apresentam assimetria tanto à esquerda (valores menores) quanto à direita (valores maiores), concentrando a maioria dos dados no centro em formato de montanha. As medidas de média, mediana e moda também são muito próximas umas das outras. No entanto, enquanto as amostras de 10% e 5% são mesocúrticas (com uma distribuição equilibrada, nem muito achatada nem muito alongada), a amostra perfeita é leptocúrtica, mais alongada.

Adicionalmente, os gráficos revelam que o valor mínimo é 0 no eixo x, e o valor máximo atinge aproximadamente 800 para as notas, confirmado também pelas tabelas. Mesmo com a assimetria à direita (valores maiores), o terceiro quartil ou Upper Quartile indica que 75% dos valores são menores que 574,3, com apenas 25% acima desse valor, e um máximo de 820,8, usando como referencia o gráfico da 10P. Isso sugere que as notas estão concentradas abaixo de 600, o que é preocupante para uma prova com pontuação máxima de 1000.

Ao observarmos os gráficos de boxplot, encontramos semelhanças com a análise anterior, destacando a presença de *outliers* tanto à esquerda quanto à direita, ou seja,

valores significativamente pequenos e grandes. Nas amostras p10 e p5, esses *outliers* são particularmente extremos. Considerando que a média é em torno de 523, encontramos candidatos com notas em torno de 254, assim como valores mínimos que chegam a 0. No lado oposto, temos valores altos (também *outliers*), embora em menor quantidade, utilizando aqui como exemplo a amostra p5.

Na disciplina de **Linguagens e Códigos**, percebemos uma média de 523, com 75% dos candidatos obtendo notas abaixo de 574. Esse dado revela uma dificuldade generalizada entre os candidatos, com a maioria apresentando notas abaixo do esperado. Apenas alguns candidatos, que aparecem como *outliers* à direita, conseguem notas muito mais altas.

Esse padrão sugere uma necessidade de análise mais aprofundada por parte das instituições responsáveis, especialmente considerando que grande parte dos candidatos recém-saíram do ensino médio. Esses dados podem indicar uma possível falha na preparação para conteúdos exigidos no exame, sinalizando que talvez seja necessário um foco maior nas áreas que os candidatos apresentam mais dificuldades.

Analysis Variable : NOTA_LINGUAGENS_E_CODIGOS Nota da prova de Linguagens e Códigos											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
533.3567485	63.1301125	304.3000000	778.2000000	538.9000000	53427	3985.41	531.0000000	473.9000000	11.8363764	494.8000000	576.4000000

Figura 11. Tabela da amostragem Perfeita

Analysis Variable : NOTA_LINGUAGENS_E_CODIGOS Nota da prova de Linguagens e Códigos											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
523.5163208	73.6199377	0	791.2000000	527.3000000	104799	5419.90	542.9000000	791.2000000	14.0625869	476.7000000	574.6000000

Figura 12. Tabela da amostragem P5

Analysis Variable : NOTA_LINGUAGENS_E_CODIGOS Nota da prova de Linguagens e Códigos											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
523.0247884	73.7114304	0	820.8000000	526.6000000	209598	5433.37	547.5000000	820.8000000	14.0932958	476.1000000	574.3000000

Figura 13. Tabela da amostragem P10

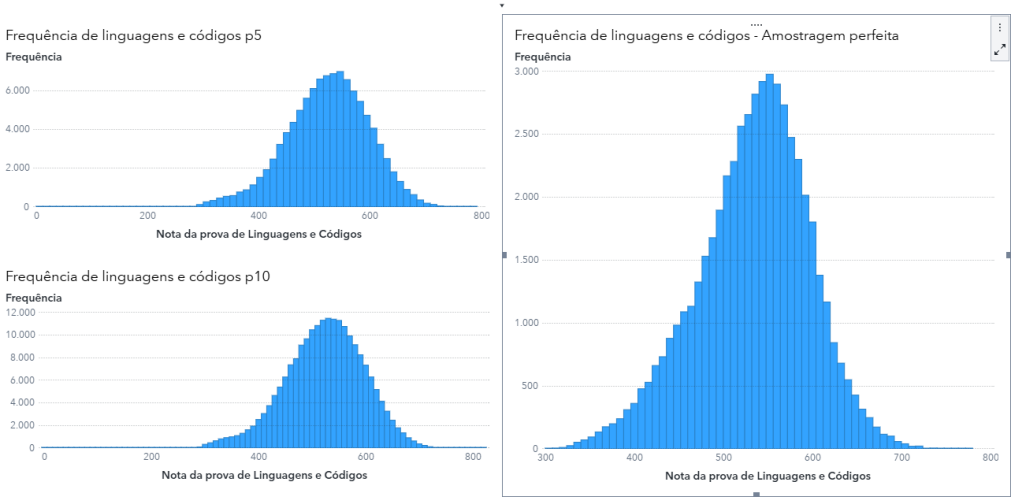

Figura 14. Histograma da de Linguagens e Códigos

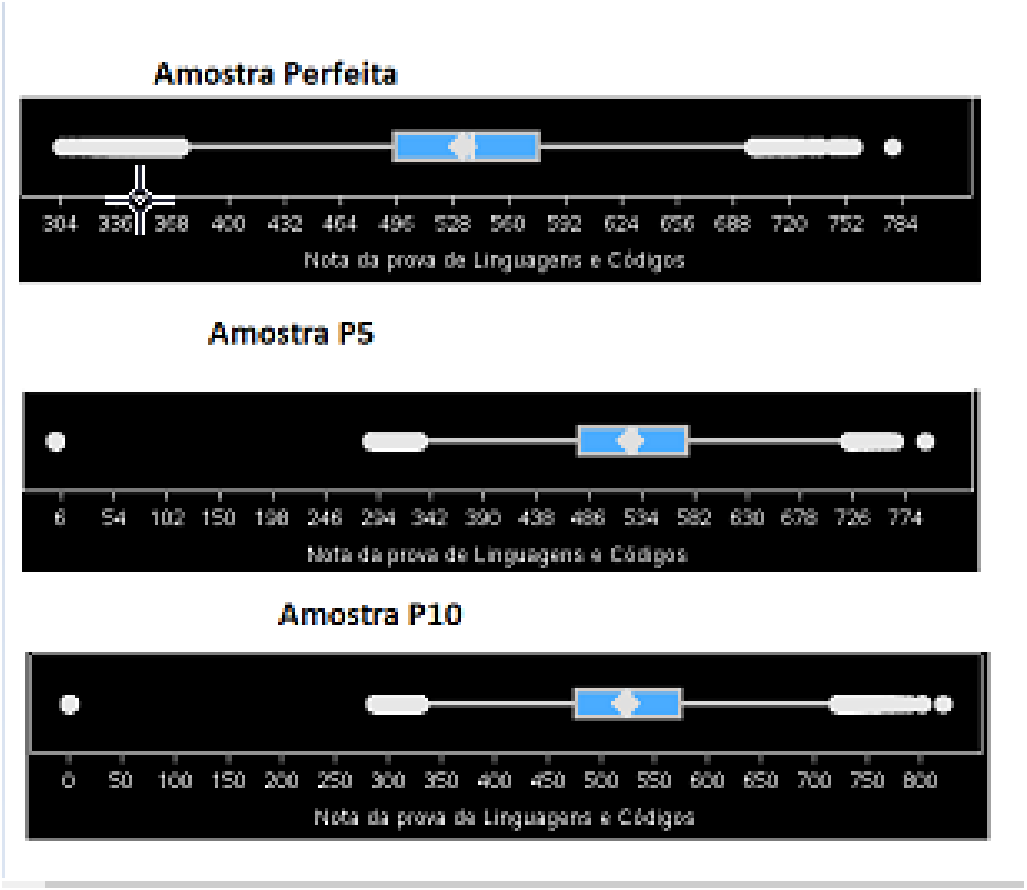

Figura 15. Box-Plot de Linguagens e Códigos

Análise da variável de Ciências Humanas

Ao analisarmos os três gráficos das notas de Ciências Humanas, percebemos que todos compartilham uma forma semelhante de crescimento e diminuição, lembrando o formato de uma montanha. No entanto, os gráficos das amostras P10 e P5 exibem uma assimetria positiva, com uma maior concentração de dados à direita. Curiosamente, essa assimetria à direita não apresenta uma cauda visivel, mas uma simetria relativa, com os dados concentrados em valores altos.

Por outro lado, o gráfico da amostra perfeita exibe uma distribuição simétrica, com dados bem equilibrados entre a esquerda e a direita.

Assim como nas variáveis quantitativas analisadas anteriormente, as amostras P5 e P10 mostram características semelhantes, o que é claramente visível nos gráficos. Em ambos, a maioria dos dados começa em torno do valor 400 e se concentra à direita. Essa tendência é evidenciada pelo quartil inferior (Lower Quartile), que indica que 25% dos dados estão abaixo de 475,40 enquanto os 75% restantes estão acima desse valor (baseando-se no gráfico da amostra P5).

Esses gráficos, assim como os da variável "Linguagens e Códigos"analisada anteriormente, apresentam um padrão de assimetria. As amostras P5 e P10 são mesocúrticas, com uma distribuição equilibrada (nem achatada nem alongada), enquanto a amostra perfeita é leptocúrtica, mais alongada.

Ao analisarmos os gráficos de boxplot das amostras p5 e p10, percebemos uma semelhança nos resultados. Observamos *outliers* tanto à esquerda (valores baixos) quanto à direita (valores altos). No entanto, é interessante notar que a quantidade de *outliers* à direita é significativamente maior, o que indica que há um grupo de candidatos com desempenho acima da média – um dado positivo.

Apesar disso, o cenário geral ainda é preocupante, já que cerca de 75% dos candidatos apresentam notas abaixo de 590,5 na amostra p10 e abaixo de 590,7 na amostra p5. Esses valores sugerem que o conhecimento da maioria dos candidatos na disciplina de **Ciências Humanas** é limitado.

Essa análise levanta um alerta sobre a preparação dos estudantes, pois, com três quartos dos candidatos obtendo notas abaixo de 590, há indícios de que a disciplina não está sendo suficientemente dominada pela maioria.

Analysis Variable : NOTA_CIENCIAS_HUMANAS Nota da prova de Ciências Humanas											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
540.3245063	68.1828195	299.9000000	804.1000000	546.0000000	53427	4648.90	561.1000000	504.2000000	12.6188649	497.7000000	587.9000000

Figura 16. Tabela da amostragem Perfeita

Analysis Variable : NOTA_CIENCIAS_HUMANAS Nota da prova de Ciências Humanas											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
530.2772030	86.5900375	0	823.0000000	536.2000000	104799	7497.83	0	823.0000000	16.3292023	475.4000000	590.7000000

Figura 17. Tabela da amostragem P5

Analysis Variable : NOTA_CIENCIAS_HUMANAS Nota da prova de Ciências Humanas											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
529.7618408	86.8029216	0	823.0000000	535.8000000	209598	7534.75	0	823.0000000	16.3852726	474.7000000	590.5000000

Figura 18. Tabela da amostragem P10

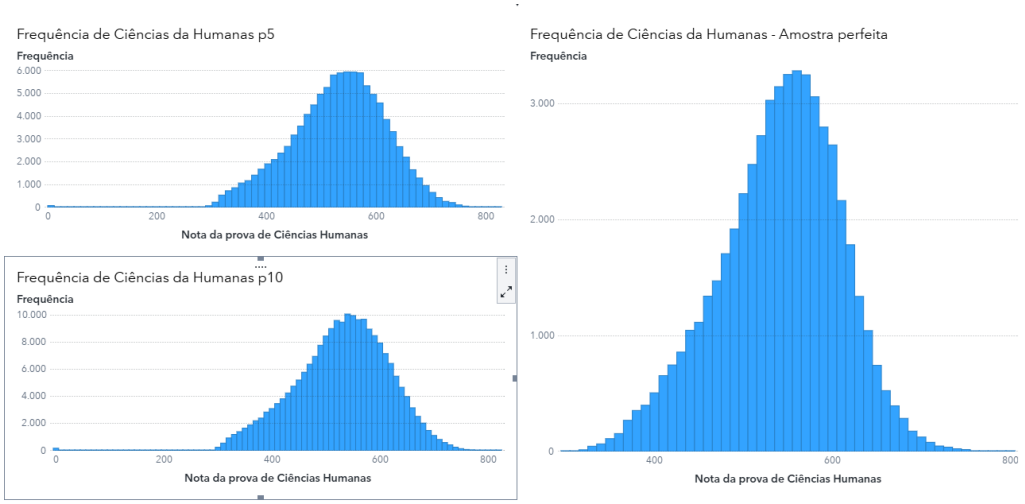

Figura 19. Histograma da Ciências Humanas

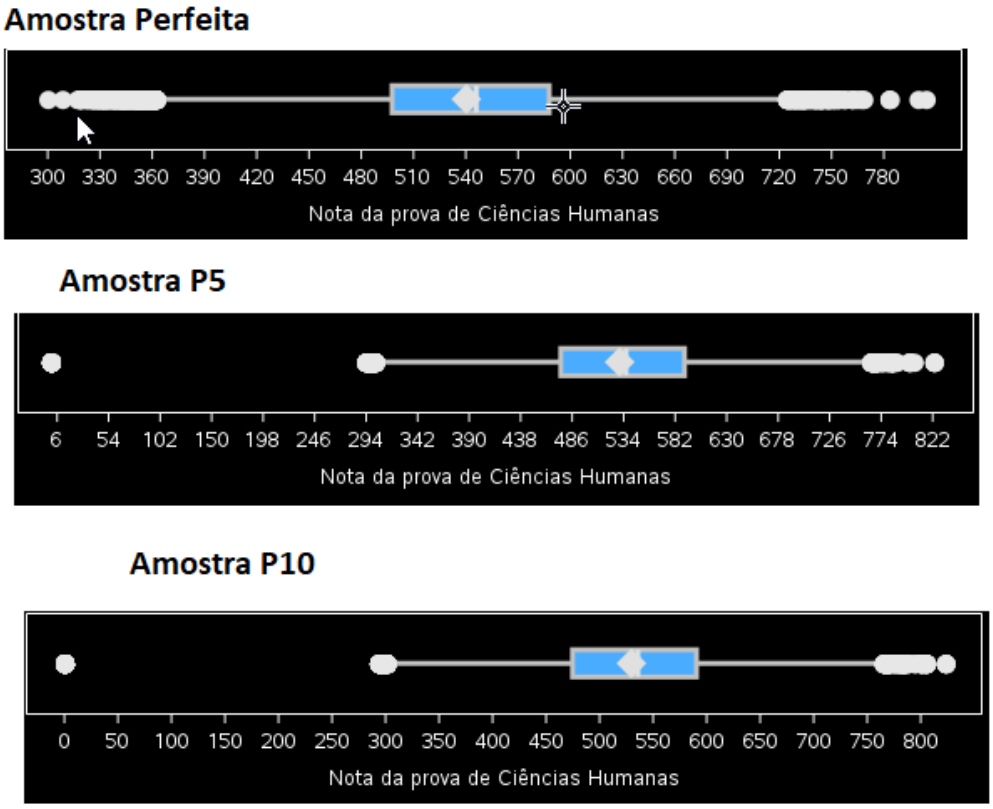

Figura 20. Box-Plot Ciências Humanas

Análise da variável de Ciências da Natureza

Ao analisarmos os três gráficos das notas de Ciências da Natureza, percebemos que as amostras P10 e P5 compartilham um formato semelhante de crescimento e queda, lembrando uma montanha. Esses gráficos também apresentam uma assimetria à direita, com uma cauda prolongada nesse sentido, o que indica que a distribuição decresce à medida que os valores aumentam. Observamos que a maior parte dos registros está concentrada acima de 443,70, com 75% dos valores até esse ponto e apenas 25% acima de 554,30 (referência com base na amostra P5). Essa observação inicial sugere que os candidatos não obtiveram notas muito altas nessa disciplina.

Outro ponto interessante é uma pequena barra visível nos valores 0 do eixo x, indicando que alguns candidatos obtiveram nota zero. Isso faz do zero a moda da distribuição, pois, como estamos tratando de valores quantitativos contínuos, o número que mais tende a se repetir é um extremo (zero ou valor máximo), em vez de valores intermediários, que geralmente aparecem em apenas uma ocorrência (como 371,50 ou 371,55). Essa característica é coerente com o desvio padrão elevado nas amostras P5 e P10, cerca de 86,62, o que indica uma grande dispersão em relação à média.

O gráfico da amostra perfeita, por outro lado, apresenta uma distribuição bem diferente, com uma forma leptocúrtica, ou seja, mais alongada e com uma concentração maior ao redor da média. Nesse caso, o desvio padrão é de 61,06, bem menor que o das outras amostras, indicando uma menor dispersão em torno da média. Esse gráfico é quase simétrico, com moda, média e mediana próximas (moda média mediana), refletindo uma

distribuição equilibrada.

Ao observarmos mais de perto, notamos em todos os gráficos um pequeno pico, seguido por um vale e logo depois um segundo pico maior, como um "morrinho"ao lado de um morro principal. Esse padrão indica que...

Ao analisarmos os gráficos de boxplot das amostras p5 e p10, percebemos uma semelhança nos resultados. Observamos *outliers* tanto à esquerda (valores baixos) quanto à direita (valores altos). No entanto, é interessante notar que a quantidade de *outliers* à direita é significativamente maior, o que indica que há um grupo de candidatos com desempenho acima da média – um dado positivo.

Apesar disso, o cenário geral ainda é preocupante, já que cerca de 75% dos candidatos apresentam notas abaixo de 553.90 na amostra p10 e abaixo de 554.30 na amostra p5. Esses valores sugerem que o conhecimento da maioria dos candidatos na disciplina de **Ciências da Natureza** é limitado.

Analysis Variable : NOTA_CIENCIAS_DA_NATUREZA Nota da prova de Ciências da Natureza											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
523.0410317	61.0692397	323.3000000	784.4000000	529.8000000	53427	3729.45	547.8000000	461.1000000	11.6758029	483.5000000	565.2000000

Figura 21. Tabela da amostragem Perfeita

Analysis Variable : NOTA_CIENCIAS_DA_NATUREZA Nota da prova de Ciências da Natureza											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
499.9506054	87.2150870	0	868.4000000	497.1000000	104799	7606.47	0	868.4000000	17.4447407	443.7000000	554.3000000

Figura 22. Tabela da amostragem P5

Analysis Variable : NOTA_CIENCIAS_DA_NATUREZA Nota da prova de Ciências da Natureza											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
499.3497447	86.6245188	0	868.4000000	496.1000000	209598	7503.81	0	868.4000000	17.3474643	443.0000000	553.9000000

Figura 23. Tabela da amostragem P10

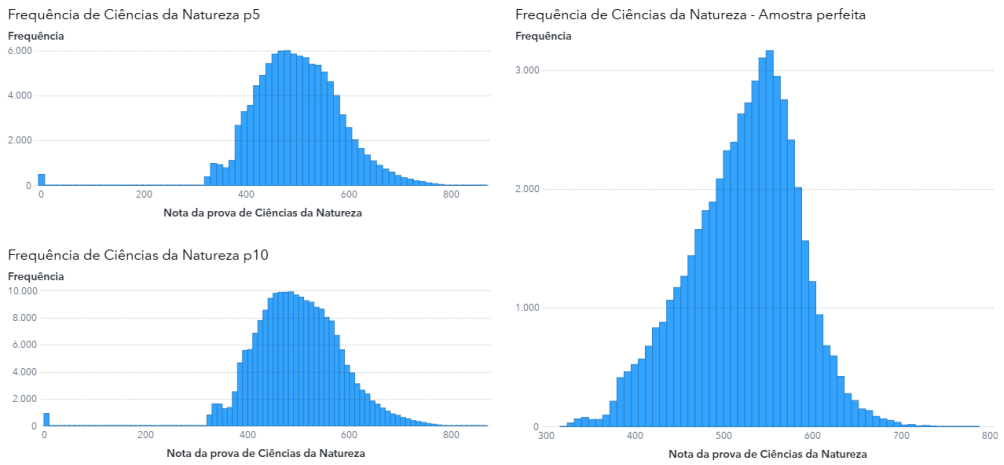

Figura 24. Histograma daCiências da Natureza

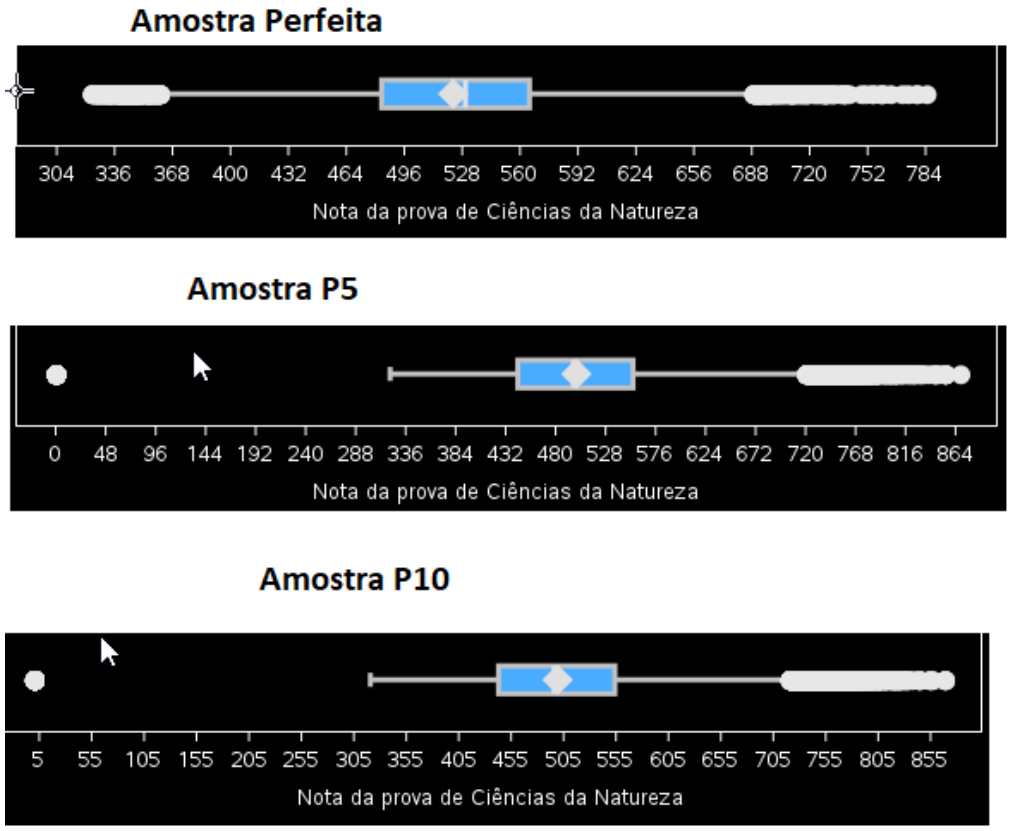

Figura 25. Box-Plor Ciências da Natureza

Análise da variável de Redação

Ao analisarmos a variável Redação, percebemos que seus gráficos diferem significativamente dos gerados por outras variáveis, apresentando barras espaçadas, quase como um gráfico de colunas discretas. Desde o início, nota-se que a distribuição é irregular, com um alto desvio padrão em todos os gráficos, especialmente nas amostras P5 e P10, que exibem desvios padrão em torno de 177 — um valor relativamente alto. Essa variação elevada reflete a presença de notas muito baixas, como 40, e notas muito altas, chegando a 1000, tudo em uma mesma análise (tomando o P10 como referência).

Observa-se que os gráficos P5 e P10 mostram um aumento nas barras da esquerda para a direita. Esse padrão se confirma ao analisarmos o primeiro quartil: 75% dos candidatos obtiveram notas acima de 520, com uma estabilidade até o valor de 800, a partir do qual a frequência cai drasticamente. Apenas 25% dos candidatos obtiveram notas superiores a 800. Ainda assim, esses dados não são homogêneos, como indica o alto coeficiente de variação: quanto maior esse valor, menos homogênea é a distribuição. Essa característica é visível na oscilação das barras, que sobem e descem de forma razoavelmente padronizada, mas ainda variada.

Já o gráfico da amostra perfeita exibe uma diferença parcial em relação aos outros. Primeiramente, possui um número de registros menor que as demais amostras, com valores máximos e mínimos de 880 e 300, respectivamente. O desvio padrão também é bem menor, indicando uma distribuição mais próxima da média e, portanto, mais homogênea

(menor coeficiente de variação). Nesse caso, cerca de 75% dos candidatos obtiveram notas abaixo de 600 e 25% abaixo de 520.

Por fim, ao observarmos os gráficos de boxplot, percebemos uma diferença interessante em relação à maioria das análises anteriores. Nas amostras p5 e p10, os *outliers* estão concentrados apenas à esquerda, ou seja, em notas significativamente abaixo da média. Esses *outliers* refletem candidatos com notas mínimas tão baixas quanto 40. Embora existam valores altos, como 1000, esses não são considerados anormais, pois cerca de 75% dos candidatos têm notas abaixo de 800 – não tão distantes desse valor máximo.

Na **amostra perfeita**, por outro lado, vemos *outliers* distribuídos em ambos os lados do gráfico, com uma média em 542,55. Esse padrão sugere uma dispersão mais ampla dos dados nessa amostra, onde há tanto notas muito baixas quanto muito altas.

Em relação à **Redação**, essa é uma das disciplinas mais temidas pelos candidatos. Ela possui caráter eliminatório em alguns casos e exige um conjunto de conhecimentos específicos para que o candidato produza um texto bem avaliado. A distribuição das notas de redação tem um comportamento atípico, lembrando oscilações de batidas de música, com variações bruscas. Observamos que, à medida que as notas aumentam, a frequência de candidatos diminui de forma acentuada. Esse formato gráfico sugere que a maior parte dos candidatos se concentra em notas médias ou baixas, com poucos alcançando as notas mais altas. Isso resulta em uma queda acentuada nas frequências das notas mais altas, indicando o nível de dificuldade e seletividade que a disciplina de Redação impõe aos candidatos.

Analysis Variable : NOTA_REDACAO Nota da prova de redação											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
542.5537650	72.2405312	300.0000000	880.0000000	540.0000000	53427	5218.69	560.0000000	580.0000000	13.3149074	500.0000000	600.0000000

Figura 26. Tabela da amostragem Perfeita

Analysis Variable : NOTA_REDACAO Nota da prova de redação											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
647.5603775	177.3478884	40.0000000	1000.00	640.0000000	104799	31452.27	560.0000000	960.0000000	27.3870815	520.0000000	800.0000000

Figura 27. Tabela da amostragem P5

Analysis Variable : NOTA_REDACAO Nota da prova de redação											
Mean	Std Dev	Minimum	Maximum	Median	N	Variance	Mode	Range	Coeff of Variation	Lower Quartile	Upper Quartile
646.8589395	177.6293436	40.0000000	1000.00	640.0000000	209598	31552.18	560.0000000	960.0000000	27.4602905	520.0000000	800.0000000

Figura 28. Tabela da amostragem P10

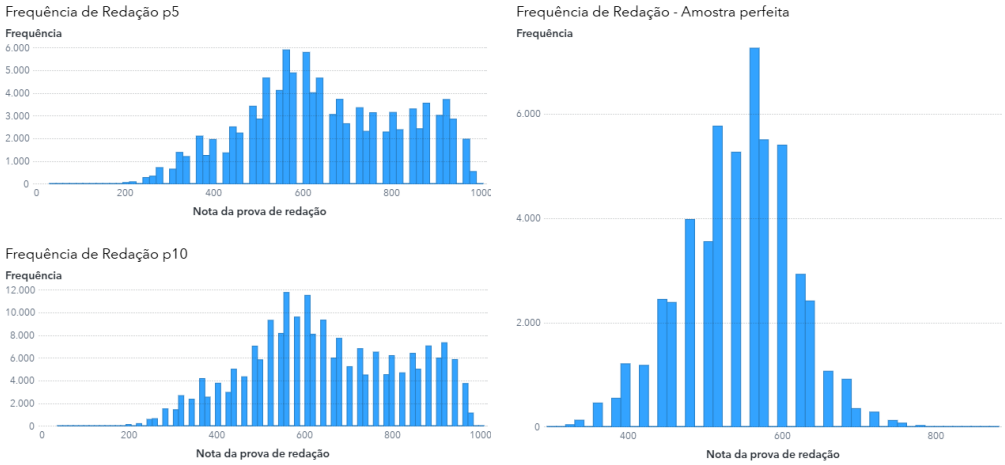

Figura 29. Histograma da Redação

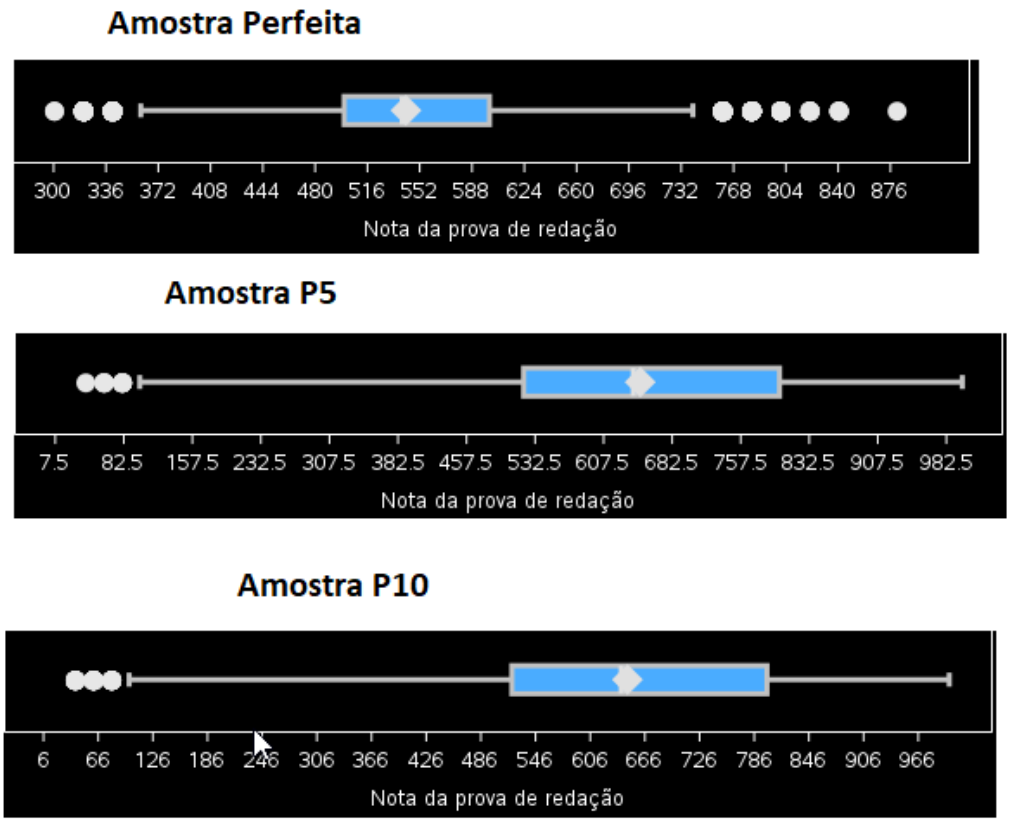

Figura 30. Box-plot da Redação

4. Conclusão

Este relatório de análise das notas do ENEM nos dá um panorama geral sobre o desempenho dos candidatos e pode ajudar as instituições a entender melhor como os estudantes estão se saindo. Em termos gerais, vemos que a maioria dos participantes tem cerca de 20 anos, com alguns casos mais raros de candidatos que chegam até os 70 anos.

No desempenho das disciplinas, a maioria não alcança notas altas. A **Redação** se destaca um pouco, sendo a única matéria onde cerca de 75% dos candidatos ficam abaixo de 800 pontos, um avanço em relação às demais, nas quais 75% ficam abaixo de 600 pontos. Esse dado, mais visível na amostra p10, sugere que o desempenho em Redação está um pouco acima da média, mas ainda assim não de forma significativa.

Para os avaliadores, o foco está nos *outliers* à direita (valores altos), que representam aqueles candidatos com melhor desempenho, pois são eles que tendem a ser mais bem beneficiados pelo programa. Além disso, essa análise serve para que as instituições de ensino identifiquem onde os estudantes recém-formados do ensino médio estão com mais dificuldade. Isso possibilita que as escolas possam ajustar suas estratégias de ensino para resolver essas lacunas.

Analisando a distribuição dos dados em diferentes amostras, conseguimos ter uma visão mais clara de como os resultados variam. As medidas de resumo, como a média e a mediana, também influenciam a forma como esses dados são apresentados nos gráficos. Essa análise foi feita usando a ferramenta SAS, que nos ajudou a visualizar e entender a estrutura dos dados.

Com este exercício, desenvolvemos o pensamento analítico, entendendo melhor a distribuição das notas do ENEM 2023. Ele serve como um ponto de partida importante para que as instituições possam repensar suas abordagens e ajudar a preparar melhor os candidatos para o futuro.